Abstract Title Page
Not included in page count.

Title:

Using a two-staged propensity score matching strategy and multilevel modeling to estimate treatment effects in a multisite observational study

Authors and Affiliations:

Jordan H. Rickles
University of California, Los Angeles

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Multisite, or block, randomized designs can facilitate unbiased estimation of treatment effects and effect heterogeneity across sites (Seltzer, 1994). Unfortunately, such designs are not always feasible or practical, and researchers frequently analyze data generated from a non-experimental multisite setting. Multilevel, or hierarchical, modeling (Raudenbush & Bryk, 2002) is a natural and commonly used tool for estimating treatment effects in these multisite observational studies, but the methodological challenges one faces when trying to make causal inferences—particularly about effect heterogeneity—are compounded in a multisite setting (Gitelman, 2005; Raudenbush, 2008; Sobel, 2007).

When applying the potential outcomes framework (Rubin, 1974, 2005) to a multisite setting, one must consider how unit- and site-level confounders might influence the assumption of strongly ignorable treatment assignment. This issue is gaining particular research interest as it pertains to the application of propensity score methods (Arpino & Mealli, 2011; Bellio & Gori, 2003; Kim & Seltzer, 2007; Su & Cortina, 2009; Thoemmes, 2009). Additionally, concerns about the stable-unit-treatment-value assumption (SUTVA) in multisite settings have been raised (Gitelman, 2005; Hong & Raudenbush, 2005, 2006, 2008; Raudenbush, 2008; Sobel, 2007) because of potential interference issues and variation in treatment enactment across sites.

When using propensity score matching in a multisite observational setting, within-site matching is preferred because it tries to approximate a multisite randomized design in a way that can at least control for observed unit-level confounders as well as observed and unobserved site-level confounders (Kim & Seltzer, 2007). Within-site matching, however, can be rather restrictive when there is not a large pool of available control unit matches within sites and/or treatment assignment is highly selective. The lack of available focal local control units within a given site was addressed by Stuart and Rubin (2008) in relation to a treatment implemented at one site and available controls coming from both the treatment site and non-treatment sites. Their approach prioritizes within-site, or local, matches and then applies a secondary between-site, or non-local, match for the unmatched treatment units.

This paper describes how the two-stage matching strategy implemented by Stuart and Rubin (2008) can be extended to a setting where a treatment is implemented in multiple sites. Utilizing this method can help overcome the limitations of within-site matching and facilitate estimation of treatment effect heterogeneity across sites.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The study is designed to demonstrate and test the utility of the proposed two-stage matching method compared to other analytic methods traditionally employed for multisite observational studies. More specifically, the study will address the following research questions: (1) How do different specifications of the matching method influence covariate balance? (2) How do different specifications in the matching method influence inferences about treatment effect and effect heterogeneity?

The different matching method specifications include differences in the propensity score model and whether a between-site match, within-site match, or two-stage matching process is used.

Setting:
Description of the research location.

The proposed two-stage matching method seeks to facilitate causal effect estimation in research settings complicated by the following factors: (1) random assignment is not practical or feasible; (2) assignment to the treatment condition is highly selective; (3) the assignment mechanism can vary across sites; and (4) the treatment effect can vary across units and sites. Settings where such factors are present can arise in educational research on policies or programs implemented across schools, where the policies/programs target a select population and implementation can vary across schools. Examples include school-based dropout prevention programs, tutoring or support services program, disciplinary correction programs, and differential course placement.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Previous research has looked at implementing a similar method in a single-site setting (Stuart & Rubin, 2008). This study extends this work to a multisite setting, where treatment effect heterogeneity can be examined. Additionally, this study will compliment the small set of studies that have examined how propensity score model specifications influence effect estimation in multisite settings (Arpino & Mealli, 2011; Su & Cortina, 2009; Thoemmes, 2009) by testing different propensity score models along with different matching methods.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

The proposed two-stage matching method consists of three primary phases: (1) a design phase, in which one uses a two-stage matching strategy to construct treatment and control groups that are well balanced along both unit- and site-level key pretreatment covariates; (2) an adjustment phase, in which the observed outcomes for non-local control group matches are adjusted to account for differences in the local and non-local matched control units; and (3) an analysis phase, in which one estimates average causal effects for the treated units and investigates heterogeneity in causal effects through multilevel modeling. The steps in each phase are adapted from Stuart and Rubin (2008) to address a multisite research setting.

The bulk of the proposed method occurs in the design phase. First, a model that includes both unit- (X) and site-level (S) covariates is used to estimate each unit's propensity score. For example, a two-level random-intercept-and-slope hierarchical model allows for a flexible specification of treatment assignment. Given estimated propensity scores, treatment and control units are matched in two steps using a within-caliper (0.25 sd), one-to-one matching algorithm implemented iteratively for each site. In the first stage, treatment units within site j are matched to control units within site j, with the set of within-site matched units for site j retained in a data set referred to as $M1_j$ (see top panel of Figure 1). The second stage in the matching process is to find non-local matches for treatment units who were not retained in the within-site matching stage. Treatment units in site *j* who are not part of $M1_j$ are matched to control units who are not in site *j* (denoted as j'). The set of between-site matched units for site *j* are retained in a data set referred to as $M2_j$ (see middle panel of Figure 1).

To adjust for the selection of between-site matches, a supplementary part of the two-stage matching strategy is to estimate site differences in the control groups. To do this, an additional match is conducted, where the matched control units in $M1_j$ are matched to other control units outside site *j*. The resulting matched control units are retained in a data set referred to as MC_j (see bottom panel of Figure 1).

After implementing the two-stage matching process—along with the supplemental control group match—for site *j*, the process is repeated for all remaining sites (i.e., sites *j*+1 through *J*). This iterative matching process conceptually parallels a block randomized design, where treatment and control groups are created by random assignment implemented independently within each site. The matched treatment and control units for each site are combined into one data set (M) and the matched control units for each site are combined into another data set (MC).

In the adjustment phase, the primary objective is to address the following counterfactual question for non-local control group matches: what would the observed outcome for control units in $M2_j$ have been if those units had been in site *j* instead of site j'. In using the two-stage matching strategy, matching some treatment units to control units in a non-local site may introduce bias in our treatment effect estimates if observed outcomes for control units differ across sites. We can try to adjust for this bias prior to the analysis phase, however, by estimating site-level differences and extracting those differences from each control unit's observed outcome.

To do this, we can estimate each site's adjusted average effect with a hierarchical model, using the MC data set (see Equation 1 in Table 1). Based on the model estimates from Equation 1, $\hat{u}_{0j} + \hat{u}_{1j} PS_{ij}^{gd}$ represents the expected effect of site *j* relative to other sites with similar characteristics for a control unit with a given propensity score. Thus, the expected difference between unit *i*'s outcome if the unit had resided in site *j* instead of j' can be represented by the following:

$$(\hat{u}_{0j} - \hat{u}_{0j'}) + (\hat{u}_{1j} - \hat{u}_{1j'}) PS_{ij}^{gd}.$$

For each non-local matched control unit, we can use the above expected difference from the estimated model to adjust the control unit's outcome for the counterfactual condition of residing in the local site (j) instead of the non-local site (j'):

$$\tilde{Y}(0)_{ij} = Y(0)_{ij'} + (\hat{u}_{0j} - \hat{u}_{0j'}) + (\hat{u}_{1j} - \hat{u}_{1j'}) PS_{ij'}^{gd},$$

where $Y(0)_{ij'}$ is the observed outcome for control unit *i* in site j' and $\tilde{Y}(0)_{ij}$ is the adjusted control unit outcome if the unit had been in site *j*. Given uncertainty in the model parameter estimates, an extension to this approach includes multiply imputing $\tilde{Y}(0)_{ij}$ based on independent draws of the random effects from their posterior distribution.

In the analysis phase, one utilizes the matched units in the M data set to estimate treatment effects and effect heterogeneity. Inference about the average treatment effect for the treated units (ATT), or what Gitelman (2005) referred to as the group-allocation, multilevel average (GAMA), and the degree of site-level ATT variance can be made with a two-level hierarchical model that includes a treatment indicator at level 1 (see Equation 2 in Table 1). One can then include site-level mediator and/or moderator variables at level 2 to study what factors are associated with site-level effect heterogeneity.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

I use a Monte Carlo simulation study to address the research questions and demonstrate the method. The main purpose of the simulation study is to compare performance of the proposed method under different propensity score modeling options and assignment mechanisms to more conventional matching methods. The different simulation conditions are summarized in Table 2.

Results based on 100 replications suggest that the two-stage matching method improves covariate balance compared to no matching and matching that pools treatment and control units across sites (see Figure 2). Covariate balance improvement with the two-stage method is similar to improvement from the more restrictive within-site matching method. The two-stage method retains more treatment units in the matched sample than the within-site method, however (see Figure 3). Furthermore, inferences regarding the average treatment effect (see Figure 4) and effect heterogeneity (see Figure 5) under the two-stage matching method are insensitive to the propensity score model used and similar to other matching methods.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

The simulation results indicate that the two-stage matching method balances the desire for within-site covariate balance and the desire to retain as many treatment units in the analysis as possible. Relative to more straightforward matching methods, however, the two-stage matching method does not result in greater covariate balance nor less biased effect estimation. As a result, more straightforward methods that address the nested data structure—such as within-site matching or pooled matching with a random-intercept-and-slope propensity score model—might be preferable to the more complex two-stage matching method. These conclusions are based on a finite set of data generating conditions, with a small set of important confounders at both the unit and site level and a reasonable within-site sample size for matching. Future research should examine the performance of various propensity score model and matching methods under more extreme data conditions.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770-1780.

Bellio, R., & Gori, E. (2003). Impact evaluation of job training programmes: Selection bias in multilevel models. *Journal of Applied Statistics*, *30*(8), 893.

Gitelman, A. I. (2005). Estimating causal effects from multilevel group-allocation data. *Journal of Educational and Behavioral Statistics*, *30*(4), 397 -412.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205 -224.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901-910.

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, *33*(3), 333 -362.

Kim, J., & Seltzer, M. (2007). *Causal inference in multilevel settings in which selection processes vary across schools*. CSE Technical Report 708, CRESST/University of California, Los Angeles.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*(1), 206-230.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322-331.

Seltzer, M. H. (1994). Studying variation in program success. *Evaluation Review*, *18*(3), 342 -361.

Sobel, M. E. (2007). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, *33*(2), 230-251.

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33*(3), 279-306.

Su, Y.-S., & Cortina, J. (2009). What do we gain? Combining propensity score methods and multilevel modeling. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450058

Thoemmes, F. (2009). *The use of propensity scores with clustered data: A simulation study* (PhD Dissertation). Arizona State University, Arizona, United States. Retrieved from http://proquest.umi.com/pqdlink?did=1895391031&Fmt=7&clientId=1564&RQT=309&VName=PQD

Appendix B. Tables and Figures

Not included in page count.

Table 1. Sample multilevel models used in the proposed two-stage matching strategy

Model	Equation
(1)	Adjustment phase school effects model:

Level 1: $Y(0)_{ij} = \beta_{0j} + \beta_{1j} PS_{ij}^{gd} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2),$

Level 2: $\beta_{0j} = \gamma_{00} + \boldsymbol{\gamma}_{\mathbf{01}} \mathbf{S}_j + u_{0j}, \quad u_{0j} \sim N(0, \tau_0)$

$\beta_{1j} = \gamma_{10} + \boldsymbol{\gamma}_{\mathbf{11}} \mathbf{S}_j + u_{1j}, \quad u_{1j} \sim N(0, \tau_1),$

where PS_{ij}^{gd} is the grand-mean centered estimated propensity score (in log-odds) for control unit *i* in site *j* and $\mathbf{S}_j$ is a vector of site-level key characteristics for site *j*. β_{0j} is the expected outcome for a control unit in site *j* who has a propensity score at the grand-mean, and resides within a given site cluster. Similarly, β_{1j} is the expected linear relationship between a unit's propensity score and the outcome at site *j*. Site-level differences in expected outcomes for the average control group unit are captured by u_{0j} and site-level differences in the expected relationship between the propensity score and outcome measure are captured by u_{1j}. Estimates of these two random effects are empirical Bayes estimates of site effects, conditional on the site characteristics. The random effects are assumed to have a multivariate normal distribution with means zero, variance τ_0 and τ_1 respectively, and covariance captured by τ_{10} (not shown in equation).

(2) Analysis phase treatment effect unconditional model:

Level 1: $Y_{ij} = D_{ij} Y(1)_{ij} + (1 - D_{ij}) \tilde{Y}(0)_{ij} = \beta_{0j} + \beta_{1j} D_{ij} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2),$

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_0),$

$\beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_1).$

where Y_{ij} is the observed outcome for unit *i* in site *j*, with treatment assignment *D* and adjusted outcome for control units $\tilde{Y}(0)$. From this model, γ_{10} represents the estimated GAMA, or the grand-mean ATT. Each site-level ATT is captured by β_{1j} and the degree to which site-level ATTs vary around the grand-mean ATT is captured by τ_1.

Table 2. Conditions tested in the Monte Carlo simulation study.

Condition Type	**Specifications**
Site-level sample size	• J=50 • J=100
Within-site sample size	• $n_j \sim N(100,5)$ • $n_j \sim N(200,10)$
Treatment assignment mechanism	• Random assignment • Selection on unit-level observables • Selection on unit- and site-level observables • Selection on unit-level observables and site-level observables and unobservables • Selection on unit- and site-level observables and unobservables
Propensity score model	• Single-level logistic regression model • Two-level random intercept (RI) logistic regression model • Two-level random intercept and slope (RIS) logistic regression model
Matching method	• Allow matches within and between sites (pooled matching) • Restrict matching to within-site • Two-stage matching method

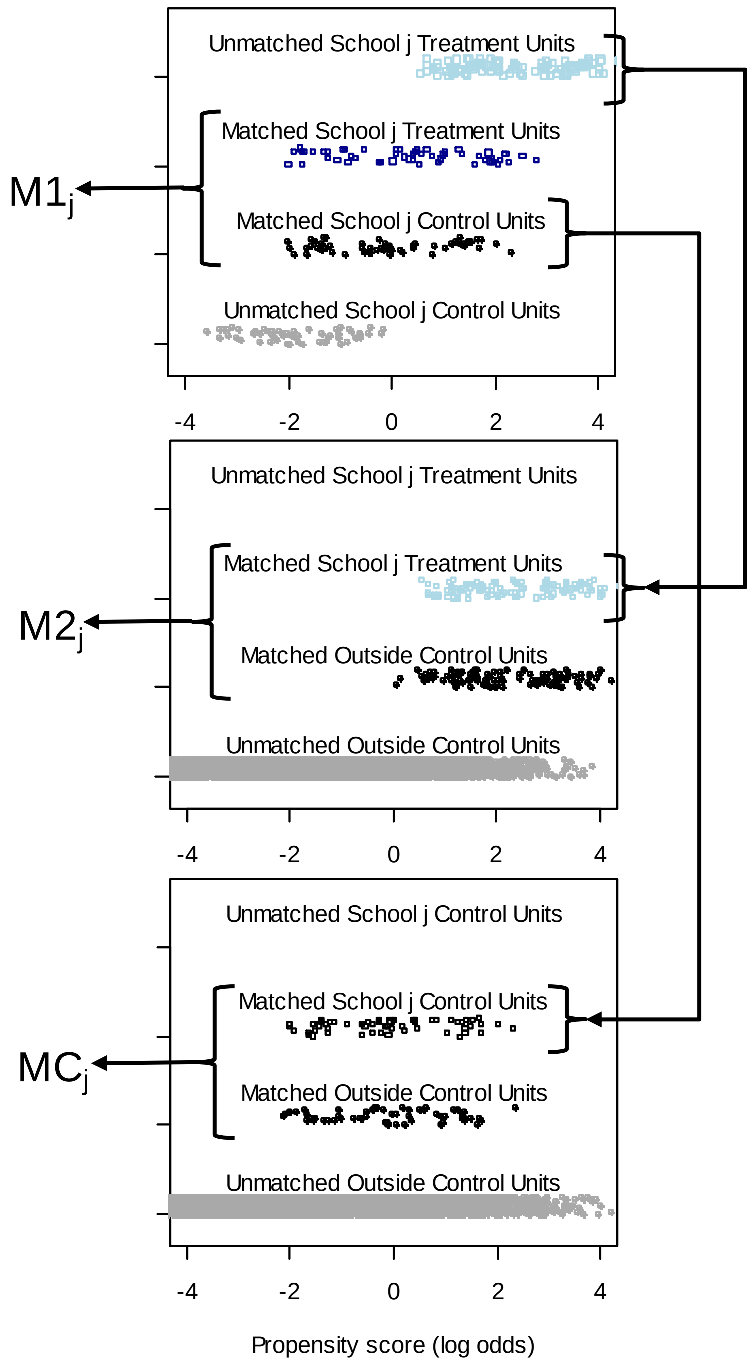

Figure 1. Jitter plots illustrating the two-stage matching strategy for one hypothetical school. Results from within-school matching (top panel) produce the $M1_j$ data set. Results from the between-school matching (middle panel) produce the $M2_j$ data set. Results from the between-school control group matching (bottom panel) produce the MC_j data set.

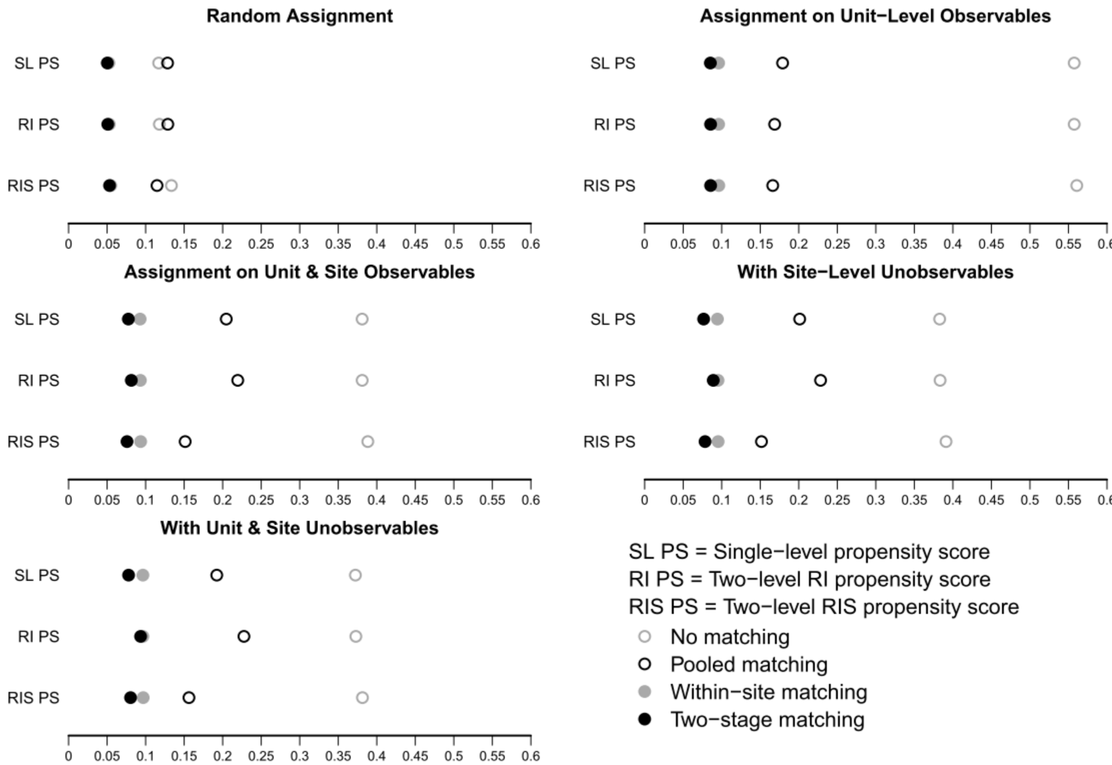

Figure 2. Group standardized bias in predicted propensity score averaged across 100 Monte Carlo replications, by simulation condition (J = 50, n_j = 200).

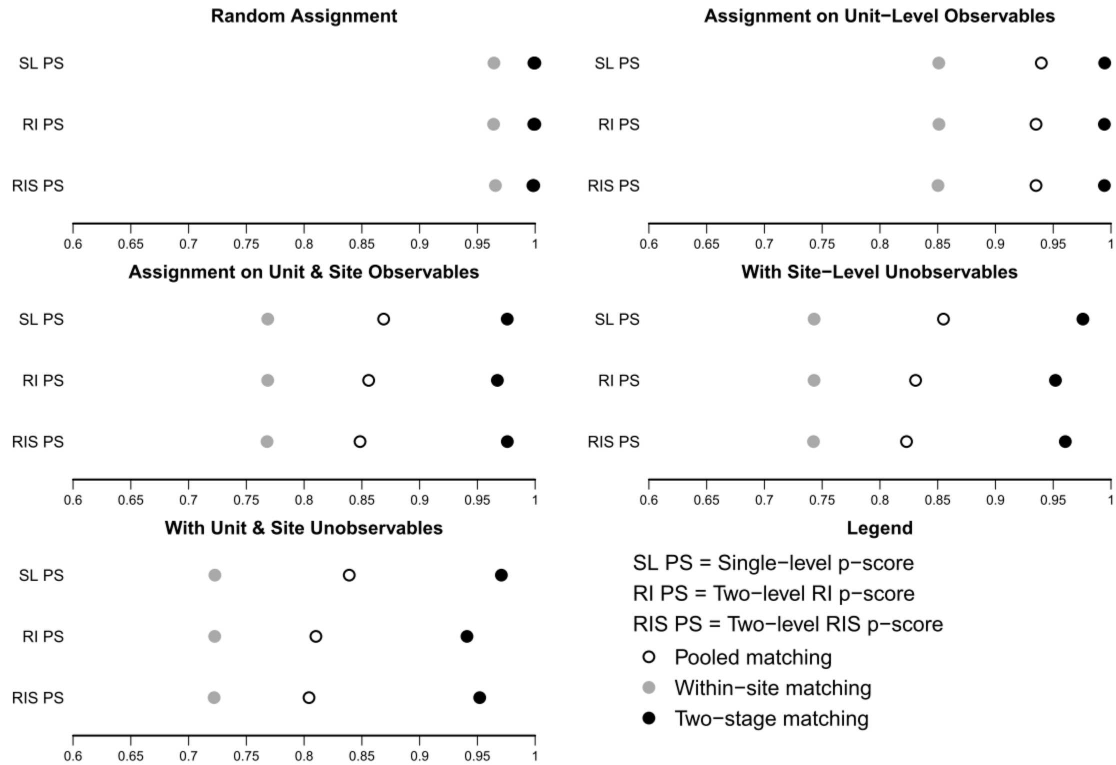

Figure 3. Mean proportion of treatment units matched across 100 Monte Carlo replications, by simulation condition (J = 50, n_j = 200).

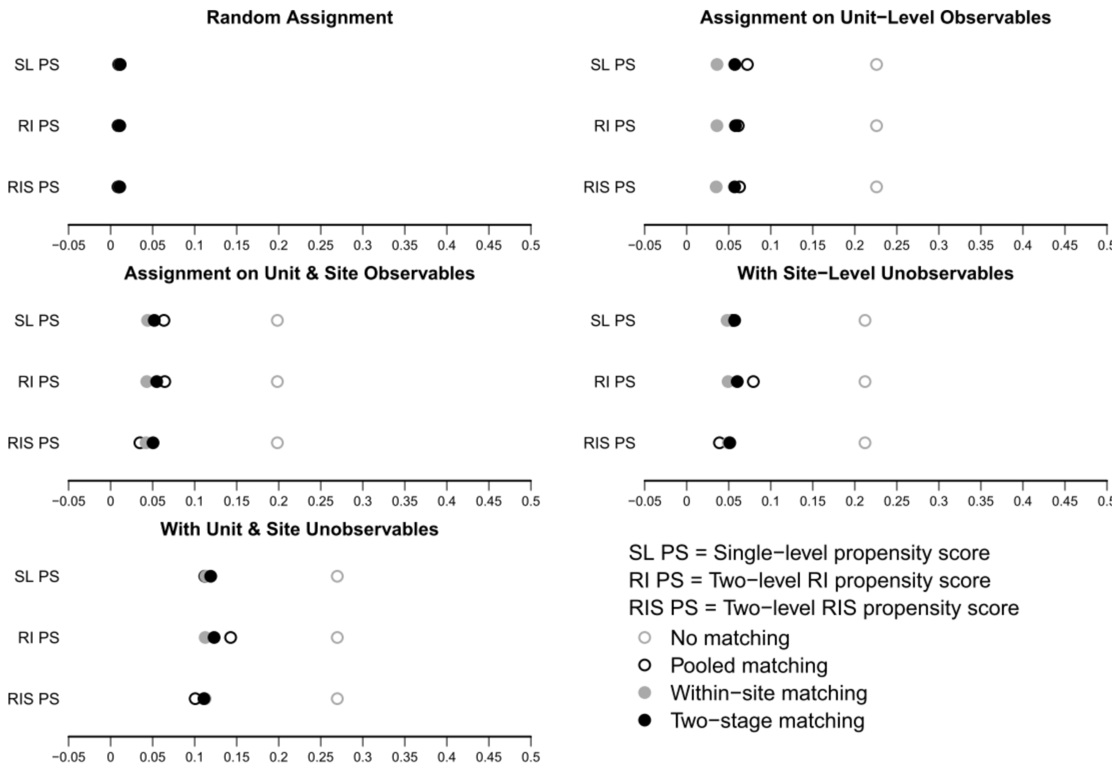

Figure 4. Mean bias in the grand-mean average treatment effect across 100 Monte Carlo replications, by simulation condition (J = 50, n_j = 200).

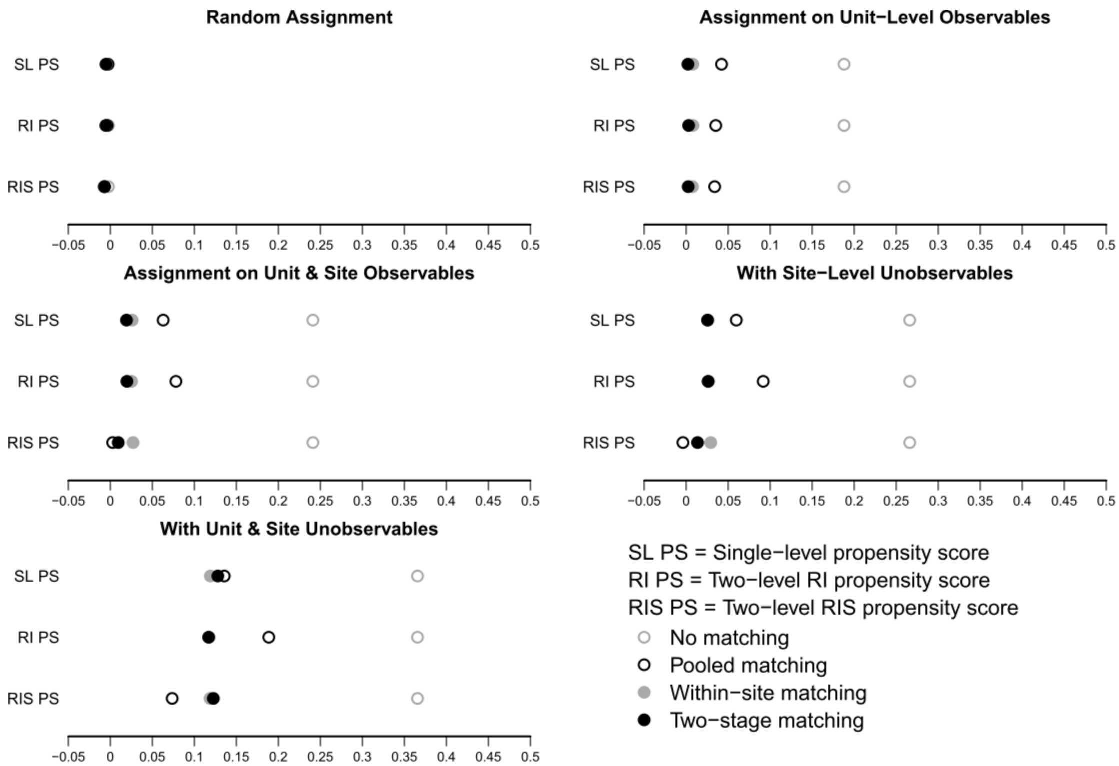

Figure 5. Mean bias in between-site average treatment effect variance across 100 Monte Carlo replications, by simulation condition (J = 50, n_j = 200).